(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2004年7月15日 (15.07.2004)

PCT

(10) 国際公開番号
WO 2004/059557 A1

(51) 国際特許分類[7]: G06F 19/00

(21) 国際出願番号: PCT/JP2003/016982

(22) 国際出願日: 2003年12月26日 (26.12.2003)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2002-377704
2002年12月26日 (26.12.2002) JP
特願2003-406776
2003年12月5日 (05.12.2003) JP

(71) 出願人 (米国を除く全ての指定国について): 独立行政法人産業技術総合研究所 (NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY) [JP/JP]; 〒100-8921 東京都 千代田区 霞が関1丁目3番1号 Tokyo (JP).

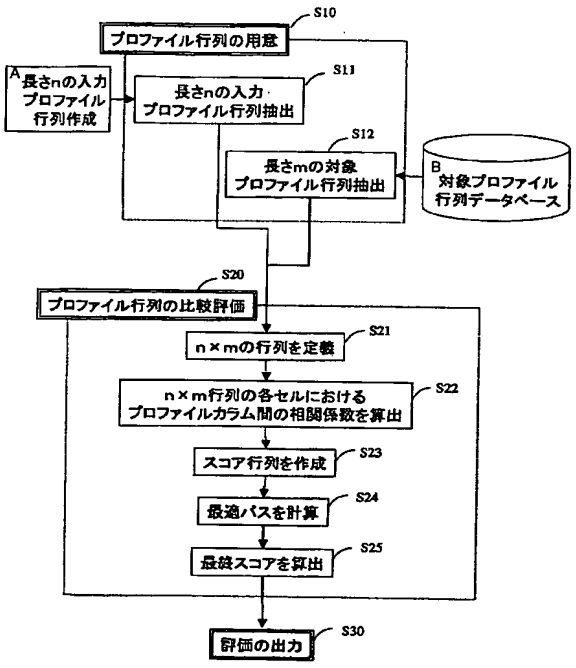
(72) 発明者; および
(75) 発明者/出願人 (米国についてのみ): 富井 健太郎 (TOMII,Kentaro) [JP/JP]; 〒135-0064 東京都 江東区 青海2-41-6 独立行政法人産業技術総合研究所内 Tokyo (JP).

(81) 指定国 (国内): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM,

[続葉有]

(54) Title: SYSTEM FOR PREDICTING THREE-DIMENSIONAL STRUCTURE OF PROTEIN

(54) 発明の名称: タンパク質立体構造予測システム



S10...PREPARE PROFILE MATRIX
A...PREPARE INPUT PROFILE MATRIX OF LENGTH OF n
S11...EXTRACT INPUT PROFILE MATRIX OF LENGTH n
S12...EXTRACT OBJECT PROFILE MATRIX OF LENGTH m
B...OBJECT PROFILE MATRIX DATABASE
S20...COMPARE AND EVALUATE PROFILE MATRIX
S21...DEFINE n×m MATRIX
S22...CALCULATE COEFFICIENT OF CORRELATION BETWEEN PROFILE COLUMNS IN EACH CELL OF n×m MATRIX
S23...MAKE SCORE MATRIX
S24...CALCULATE OPTIMUM PATH
S25...CALCULATE FINAL SCORE
S30...OUTPUT EVALUATION

(57) Abstract: A system for evaluating the similarity between protein profile matrices, preferably usable for prediction of the three-dimensional structure of a protein. A profile matrix is composed of profile columns provided with the appearance probabilities of amino acids at the positions of the amino acid residues in a multiple alignment in which the amino acid sequences of relevant proteins are multiply arranged. The similarity evaluating system comprises (a) means for preparing two matrices, an input profile matrix and an object profile matrix, (b) means for calculating the coefficient of correlation between a profile column of the input profile matrix and a profile column of the object profile matrix for all or a part of the combinations of the profile columns, and (c) means for making a score matrix composed of the coefficients of correlation.

(57) 要約: タンパク質の立体構造予測に好適に使用できる、タンパク質プロファイル行列間の類似性評価システムを提供する。本発明は、タンパク質プロファイル行列間の類似性を評価するシステムであって、プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、(a) 入力プロファイル行列と、対象プロファイル行列の2つのプロファイル行列を用意する手段と、(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、(c) 前記相関係数からなるスコア行列を作成する手段とを含む。

WO 2004/059557 A1

HR, HU, ID, IL, IN, IS, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国 (広域): ARIPO 特許 (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア特許 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ特許 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI 特許 (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書

２文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明細書

タンパク質立体構造予測システム

技術分野

本発明は、タンパク質プロファイル行列間の類似性を評価するシステムに関するものであり、より詳しくは、タンパク質の立体構造予測に好適に使用されるタンパク質プロファイル行列間の類似性の評価システムに関する。

背景技術

自然界にあるタンパク質は進化の過程で選択され、特定の機能を発現するに至ったが、このタンパク質の機能はその立体構造に依存することが知られている。したがって、タンパク質の立体構造が予測できれば、その機能を予測することが可能となる。

従来、未だ何の知見も得られていないタンパク質を調べるに際し、既に立体構造が知られているタンパク質との類似性をコンピュータによって測定することにより、タンパク質の立体構造を推論ないし予測する手法が望まれていた。このような手法の１つとして、タンパク質プロファイル行列同士を比較する方法が、有力な手法として知られている（Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci (2000) Feb;9(2):232-41：非特許文献１）。

ここで、タンパク質プロファイル行列とは、関連するタンパク質（タンパク質ファミリーなど）におけるアミノ酸種の出現頻度を、そのアミノ酸残基位置毎に数値化して行列としたものである。この行列は、通常、以下の手順で作成される。すなわち、まず、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントが与えられると、マルチプルアラインメントの各アミノ酸残基位置における２０種のアミノ酸の各種類の出現数が計算される。続いて、これらの数を規格化することによって、出現確率に転換される。この時、与えられたマルチプルアラインメントに含まれるメンバー内での相互のアミノ酸配列類似性に応じた重みが考慮された上で出現数が補正され、プロファイル行列が作

成される。

ここで、マルチプルアラインメントとは、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列を、対応すると考えられるアミノ酸残基を揃えて並置したものをいう。マルチプルアラインメントは、例えば、ある一配列を入力値として、既存のプログラムであるPSI-BLAST(Altschul et al., Nucleic Acids Res. (1997) 25(17):3389-3402：非特許文献２)を用いて、配列データベースに検索をかけることや、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列の一群を入力値として、これも既存のプログラムであるCLUSTALW(Higgins D., Thompson J., Gibson T.Thompson J.D., Higgins D.G., Gibson T.J.(1994). Nucleic Acids Res. 22:4673-4680：非特許文献３)を用いることで容易に作成することができる。また、立体構造比較などの結果から作成することも可能である。

表１は、アミノ酸配列の長さ（アミノ酸残基数）がｎであるタンパク質を基準として作成されたマルチプルアラインメントを模式的に示したものである。なお、表１中、第１列目は個々のタンパク質の名称であり、第１行目の「１～ｎ」は、マルチプルアラインメントにおけるアミノ酸残基位置を示す。また、表１中のアルファベットはアミノ酸種を１文字標記したものである。

【表１】

	1	2	3	4	5	6	7	8	…	n
20807455/14-218	M	I	D	H	T	L	L	K	…	G
19551629/13-215	I	L	D	Y	T	L	L	G	…	A
16974933/15-229	L	M	D	L	T	T	L	N	…	A
16120769/20-234	L	M	D	L	T	T	L	N	…	A

表１の例では、例示されたアミノ酸残基位置のすべてにアミノ酸が配置されているが、アミノ酸残基位置に対応するアミノ酸残基がないとされた場合は、「・（ドット）」としてギャップを示すこともできる。表２は、表１で得られた長さがｎであるマルチプルアラインメントにしたがって作成されたプロファイル行列を模式的に示したものである。表２中、第１列目はアミノ酸種（ギャップを含ん

でいてもよい）であり、第１行目の「１～ｎ」は、プロファイル行列におけるアミノ酸残基位置を示す。

【表２】

AA/Pos.	1	2	3	...	n
A	0.00	0.00	0.00	...	0.71
R	0.00	0.00	0.00	...	0.00
N	0.00	0.00	0.00	...	0.00
D	0.00	0.00	0.96	...	0.00
C	0.00	0.00	0.00	...	0.00
Q	0.00	0.00	0.00	...	0.00
E	0.00	0.00	0.04	...	0.00
G	0.00	0.00	0.00	...	0.29
H	0.00	0.00	0.00	...	0.00
I	0.29	0.29	0.00	...	0.00
L	0.41	0.29	0.00	...	0.00
K	0.00	0.00	0.00	...	0.00
M	0.29	0.41	0.00	...	0.00
F	0.00	0.00	0.00	...	0.00
P	0.00	0.00	0.00	...	0.00
S	0.00	0.00	0.00	...	0.00
T	0.00	0.00	0.00	...	0.00
W	0.00	0.00	0.00	...	0.00
Y	0.00	0.00	0.00	...	0.00
V	0.01	0.01	0.00	...	0.00

プロファイル行列中の各列は、関連する複数のタンパク質における、各アミノ酸残基位置の全アミノ酸種の確率分布を表すことになる。表３は、表２に示されたプロファイル行列のうち、残基位置が「２」であるプロファイルカラムを模式的に示したものである。

【表３】

2
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.29
0.29
0.00
0.41
0.00
0.00
0.00
0.00
0.00
0.00
0.01

すなわち、表２で示されるプロファイル行列では、残基位置が２におけるアラニン（Ａ）の補正された出現確率は０．００であり、メチオニン（Ｍ）の補正された出現確率は０．４１ということになる。

従来、２つのプロファイル行列や２つのアミノ酸配列を比較及び／又は揃えるために、ダイナミックプログラミング（Needleman SB, Wunsch CD, J Mol Biol. (1970) Mar;48(3):443-53　：非特許文献４）が使用されてきた。アラインメントを作成する時に、比較される２つのアミノ酸配列や２つのプロファイル行列中のどの残基又はプロファイルカラムを対応付させるか（そこでは残基とギャップとの対応付も含まれる）決定する必要があるが、その対応付のさせ方は非常に多数考えられる。ダイナミックプログラミングは、その中から類似性スコアが最大となるような対応付を自動的に効率良く見出すアルゴリズムである。そしてまた、その対応付の結果それ自体が最終的に得たいアラインメントである。

ダイナミックプログラミングでは、通常のアミノ酸配列比較の場合は、比較さ

れる２つのアミノ酸配列、および、比較したい２つのアミノ酸配列の各々の残基ペアに対する類似性スコア（類似の度合いを示す点数）から構成されるスコア行列、プロファイル行列比較の場合は、比較される２つの代表アミノ酸配列と、比較したい２つのプロファイル行列の、各々のプロファイルカラムのペアに対する類似性スコアから構成されるスコア行列の入力を要求する。これらを入力することによって、ダイナミックプログラミングは、通常のアミノ酸配列比較の場合は、比較されるアミノ酸配列対のアラインメントとその最終スコア（類似性スコアが最大となるような最適パスを見つけることにより得られたスコア値）、プロファイル行列比較の場合は、比較される代表アミノ酸配列のアラインメント、およびその最終スコアを出力する。

したがって、ダイナミックプログラミングを使用する手法によりプロファイル行列を比較するためには、比較したい２つのプロファイル行列の類似性を精度よく評価したスコア行列を作成する必要がある。

２つのプロファイル間の類似の程度を示すスコア行列を算出する方法として、Rychlewskiらが開発した手法が知られている（Rychlewski et al. (2000), 9:p232-241）。これは、比較したいプロファイルカラムペア間の類似性スコアを、２つのプロファイルカラムを内積したものと定義づけて算出することにより、比較したい２つのプロファイル行列間のスコア行列を作成するものである。

たとえば、２つのプロファイル行列、$X = x_1 x_2 \cdots x_p \cdots x_n$（ただし、$x_p$はアミノ酸残基位置ｐにおけるプロファイルカラム）および$Y = y_1 y_2 \cdots y_q \cdots y_m$（ただし、$y_q$はアミノ酸残基位置ｑにおけるプロファイルカラム）が与えられたとき、ｎ行ｍ列のスコア行列の要素である、類似性スコアD_{qp}（プロファイルカラムx_pおよびプロファイルカラムy_q間の類似性スコア）は、下記の式によって与えられる。

【数１】

$$D_{pq} = \sum_{a}^{j} x_{pa} y_{qa}$$

[式中、x_{pa}＝プロファイルカラムx_pの要素

y_{qa}＝プロファイルカラムy_qの要素

j＝プロファイルカラムの要素数（通常２０）である。]

当該手法によれば、比較したい２つのプロファイルカラム間において、共にアミノ酸置換が激しくない出現残基種が非常に限られている場合には、内積した値も高い数値となるため、高い類似性スコアが与えられる事になる。このように出現残基種が非常に限られておりアミノ酸変異が激しくない高度に保存されている残基位置は、生体内での機能的あるいは、物理化学的要請から高度に保存された箇所と考えられ、生物学的にも重要な位置であると考えられている。上記手法では、このような領域はその類似性を精度良く評価することができると考えられる。

しかしながら、上記手法では、こうした出現残基種が限られた位置を精度良く評価することができる可能性があるものの、生物学的に重要な位置であっても、モチーフ内に存在する非保存位置や、タンパク質立体構造上露出していることが重要で極性のみが重大な意義を占める位置、あるいはその逆に埋没部分に位置し疎水性のみが保存されている位置など、アミノ酸置換が激しく生起していてもその置換パターンに共通性があると考えられるような領域に関して精度良く評価することができないという問題があった。

さらに、スコア行列の各要素（類似性スコア）の平均値は負の値である事、標準偏差もほぼ一定値である事が望まれるため、類似性スコアに対して正規化処理を施さなければならず、煩雑であるという問題もあった。

従って、プロファイル行列間において、保存領域のみならず、非保存領域の類似性も評価できる、高精度かつ簡便な手法の開発が望まれていた。

【非特許文献１】

Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci 2000 Feb;9(2):232-41

【非特許文献２】

Altschul et al., Nucleic Acids Res. (1997) 25(17):3389-3402

【非特許文献３】

Higgins D., Thompson J., Gibson T.Thompson J.D., Higgins D.G., Gibson

T.J.(1994). Nucleic Acids Res. 22:4673-4680

【非特許文献４】

Needleman SB, Wunsch CD, J Mol Biol. 1970 Mar;48(3):443-53

発明の開示

本発明は、タンパク質の立体構造を予測するための、タンパク質プロファイル行列同士の類似性を評価するシステムを提供することを目的とする。

すなわち、本発明は、次のようなタンパク質プロファイル行列間の類似性評価システム、タンパク質立体構造の予測システム、コンピュータをそれらシステムとして機能させるためのプログラム、そのプログラムを記録したコンピュータ読み取り可能な記録媒体等を提供する。

（１）タンパク質の立体構造を予測するための、タンパク質プロファイル行列間の類似性を評価するシステムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の２つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

を含むシステム。

（２） （１）記載のシステムにより作成されたスコア行列を用いることを特徴とするタンパク質立体構造の予測システム。

（３）コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロ

ファイル行列間の類似性を評価するシステムとして機能させるためのプログラムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の２つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

を含むプログラム。

（４）上記（３）記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

（５）タンパク質プロファイル行列間の類似性を評価する方法であって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価方法は、以下のステップ：

(a) 入力プロファイル行列と、対象プロファイル行列の２つのプロファイル行列を用意するステップと、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出するステップと、

(c) 前記相関係数からなるスコア行列を作成するステップと

を含む方法。

（６）前記対象プロファイル行列が、立体構造が既知である複数のタンパク質に基づいて作成されるプロファイル行列であり、前記入力プロファイル行列が、立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成されるプロファイル行列である上記（５）記載の類似性評価方法。

（７）上記（５）又は（６）で得られたスコア行列を用いることを特徴とするタンパク質立体構造の予測方法。

本発明により、タンパク質プロファイル行列間の類似性を簡便かつ精度よく評価することができる。本発明により得られたスコア行列は、タンパク質立体構造を予測するのに好適に使用される。

図面の簡単な説明

第１図は、本発明の一実施形態において使用されるハードウエア構成を示す図である。

第２図は、本発明のプロファイル行列間類似性評価システムを含む処理手順の一例を示すフローチャートである。

第３図は、本発明のプロファイル行列間類似性評価システムにおいて、各プロファイルカラムペア毎に類似性を評価し、スコア行列を作成するステップを示す図である。

第４図は、実施例１、比較例１及び比較例２において出力された予測結果の信頼度と感度とをプロットした図である。

第５図は、実施例１及び比較例３において出力された予測結果の信頼度と感度とをプロットした図である。

発明を実施するための最良の形態

以下、本発明を詳細に説明する。

１．類似度評価システム

第１図は、本発明の一実施形態において使用されるハードウエア構成を示す図である。

第１図に示すように、本発明の類似性評価システムは、CPU101、ROM102、RAM103

、入力部104、情報通信送信/受信部105、出力部106、ハードディスクドライブ(HDD)107及びCD-ROMドライブ108等を備える。

CPU101は、情報記憶手段（例えば磁気的及び／又は光学的記録媒体）に記憶されているプログラムに従って、類似性評価システム全体を制御する。そして、入力部104などから受け取った情報を出力部106に供給する。また、ネットワーク回線109を通じて受け取った情報に基づいて評価処理を実行することもできる。入力部104は、キーボードやマウス等であり、評価処理を実行する上で必要な条件又はデータを入力するときに操作される。ROM102は、本発明の類似性評価システムの動作に必要な処理を命令するプログラム等を格納する。RAM103は、類似性評価システムにおける処理を実行する上で必要なデータを一時的に格納する。

送信／受信部105は、CPU101の命令に基づいて、ネットワーク回線109等との間で情報通信（データの送受信処理）を実行するものであり、例えばモデム、ルーター等が例示される。出力部106は、入力手段104から入力されたプロファイルデータ、その他各種条件等を、CPU101からの命令に基づいて情報表示処理する（例えば表示画面、プリンタ）。CD-ROMドライブ108は、CPU101の指示に基づいて、CD-ROMに格納されている類似性評価システムを機能させるためのプログラム又はデータ等を読み出し、例えばRAM103に格納する。CD-ROMの代わりに記録媒体として書き換え可能なCD-R、CD-RWを用いることもできる。その場合には、CD-ROMドライブ108の代わりにCD-R又はCD-RW用ドライブを設ける。また、上記媒体の他に、DVD、MOとそれらの媒体を用い、それに対応するドライブを備える構成としてもよい。

コンピュータに本発明の類似性評価システムを機能させるためのプログラムは、例えばC言語等で書くことができる。従って、このソフトウエアはWindows（登録商標）95/98/2000、Linux（登録商標）、UNIX（登録商標）等の各種オペレーティングシステムで作動させることが可能である。

第２図は、本発明のプロファイル行列間類似性評価システムを含む処理手順の一例を示すフローチャートである。

第２図に示すように、本発明にかかる類似性評価システムでは、まず、比較したい２つのプロファイル行列（入力プロファイル行列と対象プロファイル行列）を用意し、続いてそれらの類似性を評価し、必要に応じて評価結果を出力する。

以下、各処理について詳細に説明する。

（ａ）プロファイル行列の用意（Ｓ１０）

プロファイル行列を用意するステップでは、比較したい２つのプロファイル行列が用意（抽出）される（Ｓ１１、Ｓ１２）。ここで、２つのプロファイル行列のうち、一方（対象プロファイル行列）は、立体構造が既知である複数のタンパク質に基づいて作成されたプロファイル行列（第２図中、長さｍ）である。他方（入力プロファイル行列）は、立体構造を予測したいタンパク質（立体構造は未知であると既知であるとを問わない）を含む複数のタンパク質に基づいて作成されたプロファイル行列（第２図中、長さｎ）であることが好ましい。

プロファイル行列の作成方法としては、上述した従来知られている方法を採用することができ、特に制限はない。たとえば、ある一配列を入力値として、既存のプログラムであるPSI-BLASTを用いて、配列データベースに検索をかけてマルチプルアラインメントを作成し、このマルチプルアラインメントに基づいてプロファイル行列を作成してもよい。また、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列の一群を入力値として、既存のプログラムであるCLUSTALWを用いてマルチプルアラインメントを作成し、当該マルチプルアラインメントに基づいてプロファイル行列を作成してもよい。また、予め作成されたマルチプルアラインメントを入力値とし、このマルチプルアラインメントに基づいて作成してもよい。

ここで、プロファイル行列は、ある代表アミノ酸配列の全配列に基づいて作成されていてもよく、また、代表配列中のモチーフ領域等、一部の領域に基づいて作成されていてもよい。また、マルチプルアラインメントを作成する際に、経験的に導出されたギャップペナルティーを導入してもよい。

また、必要に応じて、プロファイル行列として、アミノ酸種の出現頻度を、アミノ酸種のランダム出現頻度で割った行列（ＰＳＳＭ：Gribskov, M., et al., (1987) Proc. Natl. Acad. Sci. USA, 84, 4355-4358）を用いてもよい。

入力プロファイル行列は、たとえば、立体構造を予測したいタンパク質を代表アミノ酸配列として、この配列に基づいて作成することができる。また、対象プロファイル行列については、たとえば、ＳＣＯＰ(Murzin et al., J. Mol. Biol.

247(4):536-540 (1995))やＣＡＴＨ(Orengo et al., Structure 5(8):1093-1108 (1997))といったタンパク質構造分類データベースから取得したタンパク質のアミノ酸配列を代表配列とし、この配列に基づいて作成することができる。こうして得られた対象プロファイル行列は、代表配列ごとに予め作成しておき、対象プロファイル行列データベースとして保持しておくことが好ましい。

（ｂ）相関係数の算出（プロファイル行列の比較評価）（Ｓ２０）

続いて、プロファイル行列の類似性評価ステップでは、上記のステップで用意した入力プロファイル行列の各プロファイルカラムと、対象プロファイル行列の各プロファイルカラムとの間の類似性を、各カラムペア毎に評価をする。

第３図は、各プロファイルカラムペア毎に類似性を評価し、スコア行列を作成するステップを模式的に示した図である。

本発明において、プロファイルカラム間の類似性は、プロファイルカラム間の相関係数を算出することによって行う。

たとえば、入力プロファイル行列を$X = x_1 x_2 \cdots x_p \cdots x_n$（ただし、$x_p$はアミノ酸残基位置ｐにおけるプロファイルカラム）とし、対象プロファイル行列を$Y = y_1 y_2 \cdots y_q \cdots y_m$（ただし、$y_q$はアミノ酸残基位置ｑにおけるプロファイルカラム）としたときに、プロファイルカラムx_pおよびy_q間の類似性スコアc_{qp}は、下記の式によって与えられる。

【数２】

$$C_{pq}=\frac{\sum_{a}^{j}(x_{pa}-\overline{x_p})(y_{qa}-\overline{y_q})}{\sqrt{\sum_{a}^{j}(x_{pa}-\overline{x_p})^2\sum_{a}^{j}(y_{qa}-\overline{y_q})^2}}$$

[式中、x_{pa}＝プロファイルカラムx_pの要素

y_{qa}＝プロファイルカラムy_qの要素

$\overline{x_p}$＝プロファイルカラムx_pの平均値

$\overline{y_q}$＝プロファイルカラムy_qの平均値

j＝プロファイルカラムの要素数（通常２０）である。]

本発明では、プロファイルカラム間の類似性をプロファイルカラム間の相関係数によって評価する。このため、プロファイルカラム間の相関の程度によって、類似性スコアが＋１から－１の値をとることになる。たとえば、２つのプロファイルカラム中の要素間に相関がある場合、即ちアミノ酸置換パターンの傾向に類似性が有る場合には、相関係数は＋１に近い数値を取ることになる。また、２つのプロファイルカラムの各要素が互いにランダムな値を取っている場合、即ちアミノ酸置換パターンの傾向に相関が無い場合、相関係数は０になり、アミノ酸置換パターンの傾向が全く反対の場合、相関係数は－１になり、アミノ酸置換パターンの傾向性の類似－非類似を非常に自然な形で表現する事が出来る。

したがって、本発明では、アミノ酸残基の保存性が高い保存領域のような相関が高い領域では、高い類似性スコアが得られるため、保存領域の類似性を精度よく評価することができる。

また、本発明によれば、アミノ酸残基の保存性だけではなく、内積によって類似性を評価する従来の方法（Rychlewski et alら）では不可能であった領域に関する類似性評価、たとえば、モチーフ内に存在する非保存位置や、タンパク質立体構造上露出していることが重要で極性のみが重大な意義を占める位置、あるいはその逆に埋没部分に位置し疎水性のみが保存されている位置といった、激しいアミノ酸置換があるもののその置換パターンに共通性があると考えられる領域に

ついての類似性をより精度良く評価することが可能である。

例えば、あるzinc fingerモチーフを有する２つのプロファイル行列を比較した場合を考えたとする。そのモチーフは

C-[DES]-x-C-x(3)-I

と表記される。これは、1，4，8番目の残基にそれぞれC，C，Iの残基が保存されており、2番目の残基では、D又は E又は Sが出現し、3番目および、5，6，7番目の残基では保存残基が特に無いことが表されている。内積によって類似性を評価する従来の方法では、この場合、1，2，4，8番目の残基位置では、高い数値を与えるが、その他の位置では低い数値しか与えない。したがって、内積によって類似性を評価する従来の方法は、モチーフの一部については類似性を評価しているものの、モチーフ全体の類似性については精度よく評価していないということになる。

しかしながら、本発明によれば、1，2，4，8番目の残基位置に高い数値を与えるだけでなく、3，5，6，7番目の残基位置においても、保存残基が特に無いという置換パターンの類似性を評価することが可能で、これら残基位置でも高い数値を与える。したがって、本発明によれば、モチーフ全体としてのパターン情報の全てを評価することが可能となる。

なお、本発明における類似性評価システムは、モチーフ領域に限られず、立体構造を予測したいタンパク質の配列全体に適用することができる。すなわち、ギャップペナルティを導入して得られたプロファイル行列間の類似性評価にも、好適に適用することができる。

さらに、本発明によれば、スコア行列の各要素（類似性スコア）の平均値および標準偏差がほぼ一定値をとるため、類似性スコアに対する煩雑な正規化処理を施す必要がないというメリットもある。

（ｃ）スコア行列の作成

プロファイルカラム間の相関係数（類似性スコア）は、各プロファイルカラムの全部又は一部の組合せについて算出され、これに基づいてスコア行列が作成される。スコア行列は、類似性スコアが各プロファイルカラムの全組合せについて算出された場合は、入力プロファイル行列の長さを行とし、対象プロファイル行

列の長さを列とする行列であり、類似性スコアが各プロファイルカラムの一部の組合せについて算出された場合は、その組合せの数に応じた行と列を持つ行列となる。

第２図の例では、類似性スコアは各プロファイルカラムの全組合せについて算出されており、入力プロファイル行列の長さがｎ、対象プロファイル行列の長さがｍであることから、類似性スコアはｍ×ｎ個生成される（Ｓ２２）。したがって、スコア行列はｎ行ｍ列となる。スコア行列は、比較したいプロファイル行列の長さ、及び算出される類似性スコアの数に応じた行列を予め定義し（Ｓ２１）、定義された行列の各カラムに、各プロファイルカラム間の相関係数を入力することにより作成することができる（Ｓ２３）。

本発明で得られたスコア行列によって、２つのプロファイル行列の最終スコア（行列間の類似性）を精度よく算出することができる。最終スコアは既知の手法により作成することができる。たとえば、第２図の例では、比較されるプロファイル行列のそれぞれの代表アミノ酸配列と、本発明によって得られたこれらのプロファイル行列間のスコア行列を入力値として、ダイナミックプログラミングを用いて最適パスを算出する（Ｓ２４）ことによって最終スコアを求めることができる（Ｓ２５）。

以上の操作を、対象プロファイル行列データベースに保持してある対象プロファイル行列のすべてに対して行うことが好ましい。

２．タンパク質立体構造の予測（Ｓ３０）

対象プロファイル行列ごとに得られた最終スコアは、タンパク質立体構造を予測するのに好適に使用される。たとえば、以下の既知の手順にしたがって処理をされる。

(1)　入力値

まず、予測対象配列を含む入力プロファイル行列と、立体構造が既知である代表アミノ酸配列を含む対象プロファイル行列との最終スコア、および各代表配列の長さが入力される。このとき、対象プロファイル行列データベース中にＮ本の既知代表配列があれば、Ｎ個の最終スコアと配列長が入力されることになる。

(2)　最終スコアの長さ依存性の補正

予測対象配列を含む入力プロファイル行列と、各既知代表配列を含む対象プロファイル行列との最終スコアは、代表配列長に依存した関係が認められる為、次のような統計処理を行う。まず、X軸に各代表配列の長さの自然対数をとった値、Y軸に予測対象配列を含む入力プロファイル行列と各既知代表配列を含むプロファイル行列との最終スコアをプロットし、異常なはずれ値を除いて回帰直線を引く。各長さにおける最終スコアの平均値は回帰直線で表されるものとみなし、予測対象配列を含む入力プロファイル行列と各既知代表配列を含む対象プロファイル行列との最終スコアは、平均値からのずれで評価される。通常良く使用されるように、標準偏差を単位として、そのずれの度合いが測定される。

(3) ソート

平均値からのずれが（高得点側に）大きいもの程類似性が有るとみなされる。それ故、平均値からのずれが（高得点側に）大きい順にソートされ、予測構造の候補とされる。

(4) 予測構造としてのアラインメントとスコア出力

上でソートされた順に予測構造の候補として出力される。結果全てを出力するのは無意味なため、予測精度を考慮し経験的に求められた閾値以上の平均値からのずれを有する結果のみを出力する。この時、予測精度の指標として、標準偏差を単位として計算される平均値からのずれの度合いが表示される。

予測対象配列を含む入力プロファイル行列と、各既知代表配列を含む対象プロファイル行列とのアラインメントおよび最終スコアの結果は、ダイナミックプログラミングを用いて逐次計算された際のものを出力する。各既知代表配列は立体構造既知なので、このアラインメント出力が立体構造予測結果に相当する。

3．コンピュータプログラム

本発明は、コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムとして機能させるためのプログラムをも提供する。本発明のコンピュータプログラムは、以下の手段：

(a) 入力プロファイル行列と、対象プロファイル行列の２つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイ

ル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

を含むものである。

本発明のプログラムには、上記必須の手段以外に、汎用のプログラムとして通常備えられる汎用手段を含んでもよい。そのような手段としては、各種データの格納手段、情報の送受信手段、ディスプレイ、プリンター等の表示・出力手段等を挙げることができる。

4．コンピュータ用記録媒体

本発明のプログラムは、コンピュータ読み取り可能な記録媒体又はコンピュータに接続しうる記憶手段に保存することができる。本発明のプログラムを含有するコンピュータ用記録媒体又は記憶手段も本発明に含まれる。記録媒体又は記憶手段としては、磁気的媒体（フレキシブルディスク、ハードディスクなど）、光学的媒体（CD、DVDなど）、磁気光学的媒体（MO、MD）などが挙げられる。

【実施例】

以下、実施例により本発明をさらに具体的に説明する。但し、本発明はこれら実施例に限定されるものではない。

実施例1

(1) 対象プロファイル行列データベースの構築

構造分類データベースＳＣＯＰ(URL:http://scop.mrc-lmb.cam.ac.uk/scop/) release1.59 に基づく分類から、代表配列を取得した。その中から、単独ドメインを有し解像度2.5Å以内の構造データを有するタンパク質のアミノ酸配列９４８本を選択した。９４８本の代表配列各々に対してPSI-BLASTとアミノ酸配列データベース(ＮＲＤＢ:ftp://ftp.ncbi.nlm.nih.govより取得)を用いて対象プロファイル行列を構築し、対象プロファイル行列データベースを完成させた。

ここで使用した「ＮＲＤＢ」には、現在知られているほぼ大部分のタンパク質アミノ酸配列が含まれている。PSI-BLASTを使うことで、このＮＲＤＢから各代表配列に生物学的に関連あると考えられる配列を自動的に収集し、さらにプロファイル行列も作成することが出来る。

(2) 入力プロファイル行列の作成

本発明にかかるシステムによって正しい構造予測がなされているかどうかを調べるため、予測対象配列として構造が既に知られている配列、すなわち、対象プロファイル行列を作成する際に使用した上記９４８本の代表配列を使用した。入力プロファイル行列は、これらの予測対象配列を順次使用して、対象プロファイル行列の場合と同様の操作、すなわち、PSI-BLASTとアミノ酸配列データベース(NRDB)を用いて構築した。

(3) 各プロファイル行列間の比較

続いて、上記で構築された予測対象配列（本実施例では９４８本の各代表配列）を含む入力プロファイル行列と、対象プロファイル行列データベース中の対象プロファイル行列との比較が順次なされた。この際、プロファイル行列間のスコア行列の各要素（類似性スコア）は、相関係数を用いて計算された。

こうして得られたプロファイル行列間のスコア行列を入力値として、ダイナミックプログラミングによってプロファイル行列間の最終スコアとアラインメントが出力された。

各入力プロファイル行列に対して、以上の操作を対象プロファイル行列データベースに構築されたすべての対象プロファイル行列について行った。

(4) 最終処理及び結果出力

評価の出力は、既に説明した方法に従って、９４８予測について各々結果出力を行った。すなわち、入力プロファイル行列と対象プロファイル行列との各最終スコアおよび各代表配列の長さを入力し、最終スコアの長さ依存性の補正を行った。続いて、平均値からのずれが（高得点側に）大きい順にソートし、ソートされた順に予測構造の候補として出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第４図に示した。

比較例１

実施例１で取得した９４８本の代表配列を用いて、配列類似性検索として一般的に用いられているＰＳＩ－ＢＬＡＳＴを用いて構造予測を行った。すなわち、

９４８本の代表配列各々に対してPSI-BLASTとアミノ酸配列データベース(ＮＲＤＢ:ftp://ftp.ncbi.nlm.nih.govより取得)を用いて構築したプロファイル行列を入力値とし、９４８本の代表配列に対して類似性検索を行い、予測構造の候補を出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第４図に示した。

比較例２

実施例１で取得した９４８本の代表配列を用いて、配列類似性検索として一般的に用いられているＩＭＰＡＬＡ(Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., and Altschul, S. F. (1999) Bioinformatics. 015:1000-1011）を用いて構造予測を行った。すなわち、９４８本の代表配列を入力値とし、９４８本の代表配列各々に対して予め作成し構築したプロファイル行列データベース（実施例１で構築した対象プロファイル行列データベースを使用した）に対して類似性検索を行い、予測構造の候補を出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第４図に示した。

第４図から、比較例１および２の手法に比べて、信頼度０．９８以降において、本発明にかかる実施例１が常に感度で勝っていることが示される。

比較例３

プロファイル行列間のスコア行列の各要素（類似性スコア）を、内積法(Rychlewski et al. (2000), 9:p232-241)を用いて計算した以外は実施例１と同様の手法で予測構造の候補を出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第５図に示した。

実施例２

(1) 対象プロファイル行列データベースの構築

配列は、構造分類データベースSCOP(URL:http://scop.mrc-lmb.cam.ac.uk/scop/) release1.59に基づく分類から、お互いの同一残基率が40%未満であるドメイン単位の代表配列4381本を、SCOPの配列データベースであるASTRAL(http://astral.stanford.edu/)データベースから取得した。更に、タンパク質立体構造データベースPDB(URL:http://www.rcsb.org/pdb/)に登録されているが、SCOPに未登録であるものであって、ASTRALから取得した上記4381本の配列と非類似のものを下記（A）～（D）の要領で取得し、代表配列に加えた。このようにして選択されたアミノ酸配列各々に対して、下記（A）～（D）の要領でPSI-BLASTとＮＲＤＢを用いて対象プロファイル行列を構築し、対象プロファイル行列データベースを完成させた。

（A）対象プロファイル行列データベースＡの構築

2002年5月18日時点でのPDB中のアミノ酸配列をSCOPrelease1.59の分類に基づく代表配列に対してBLASTP(Altschul et al., Nucleic Acids Res. (1997) 25(17): 3389-3402：非特許文献２)をかけ、期待値が0.00001以上のものを選んだ。さらにそれらを配列のクラスタリングを行うプログラムであるblastclustにかけ、互いの同一残基率が40%未満となるように配列248本を選択した。このようにして選択された配列と、SCOPrelease1.59の分類に基づく代表配列4381本との合計4629本の配列各々に対して、PSI-BLASTと2002年5月18日時点のＮＲＤＢを用いて対象プロファイル行列を構築し、対象プロファイル行列データベースＡを完成させた。

（B）対象プロファイル行列データベースＢの構築

2002年6月23日時点でのPDBと2002年5月18日時点でのPDB中のアミノ酸配列の差分を上記（A）で作成した代表配列に対してBLASTPをかけ、期待値が0.00001以上のものを選んだ。さらにそれらをblastclustにかけ、互いの同一残基率が40%未満となるように配列49本を選択した。このようにして選択された配列と、上記（A）で作成した代表配列との合計4678本の配列各々に対して、PSI-BLASTと2002年6月17日時点のＮＲＤＢを用いて対象プロファイル行列を構築し、対象プロファイル行列データベースＢを完成させた。

（C）対象プロファイル行列データベースＣの構築

2002年7月14日時点でのPDBと2002年6月23日時点でのPDB中のアミノ酸配列の差分を上記（B）で作成した代表配列に対してBLASTPをかけ、期待値が0.00001以上のものを選んだ。さらにそれらをblastclustにかけ、互いの同一残基率が40%未満となるように配列23本を選択した。このようにして選択された配列と、上記（B）で作成した代表配列との合計4701本の配列各々に対して、PSI-BLASTと2002年7月9日時点のＮＲＤＢを用いて対象プロファイル行列を構築し、対象プロファイル行列データベースＣを完成させた。

（D）対象プロファイル行列データベースＤの構築

上記（C）で作成した代表配列の合計4701本の配列各々に対して、PSI-BLASTと2002年8月6日時点のＮＲＤＢを用いて対象プロファイル行列を構築し、対象プロファイル行列データベースＤを完成させた。

(2) 入力プロファイル行列の作成

配列は、隔年で行われる世界的規模で行われる構造予測コンテストの2002年度大会であるCASP5/CAFASP3(URL:http:// predictioncenter.llnl.gov/casp5/)において、構造認識部門（通常の配列解析手法では立体構造既知であるタンパク質と明白な配列類似性を有さないが、その構造が（実際に解かれてみると）既知立体構造との構造類似性を有する、即ち類似性検索が困難なタンパク質に関する予測部門）において出題された配列、すなわち、現在通常の配列解析手法（例えば、PSI-BLASTなど）では、立体構造既知であるタンパク質と明白な配列類似性を有さないタンパク質であり、かつ、その構造が（実際に解かれてみると）既知立体構造との構造類似性が明らかになったアミノ酸配列を用いた。具体的には、URL:http://www.cs.bgu.ac.il/~dfischer/CAFASP3/targets.htmlにおいて、下記のターゲット番号が付されたアミノ酸配列２２本を用いた。

T0130、T0132、T0134、T0135、T0136、T0138、T0146、T0147、T0148、T0156、T0157、T0159、T0162、T0168、T0170、T0172、T0173、T0174、T0186、T0187、T0191、T0193

これら22本の配列各々に対して、PSI-BLASTとＮＲＤＢを用いて入力プロファイル行列を構築し、入力プロファイル行列データベースを完成させた。

なお、ＮＲＤＢとしては、2002年5月18日時点、2002年6月17日時点、2002年7月9日時点、及び2002年8月6日時点のものの計４種類を使用し、得られた

入力プロファイル行列データベースを、それぞれ、「入力プロファイル行列データベースＡ」、「入力プロファイル行列データベースＢ」、「入力プロファイル行列データベースＣ」、及び「入力プロファイル行列データベースＤ」とした。

(3) 各プロファイル行列間の比較

続いて、上記で構築された予測対象配列を含む入力プロファイル行列データベースＡの入力プロファイル行列と、対象プロファイル行列データベースＡ中の対象プロファイル行列との比較を、実施例１の「(3)各プロファイル行列間の比較」と同様の手順で行った（比較Ａ）。

同様の操作を、入力プロファイル行列データベースＢと対象プロファイル行列データベースＢに対して、入力プロファイル行列データベースＣと対象プロファイル行列データベースＣに対して、及び、入力プロファイル行列データベースＤと対象プロファイル行列データベースＤに対して、それぞれ行った（比較Ｂ，Ｃ，Ｄ）。

(4) 最終処理及び結果出力

評価の出力は、既に説明した方法に従って22予測について各々結果出力を行った。即ち、各データベースの組合せ（比較Ａ～Ｄ）においてそれぞれ得られた、入力プロファイル行列と対象プロファイル行列との各最終スコアおよび、各代表配列の長さを入力し、最終スコアの長さ依存性を補正した。続いて平均値からのずれが、（高得点側に）大きい順にソートし、ソートされた順に上位10個までを予測構造の候補として22本の配列各々に対して出力した（出力Ａ～Ｄ）。

こうして出力された予測構造の候補と、コンテストの予測構造投稿期間の後に公開された実験により解かれた立体構造とを比較することで、予測結果の正確さが測定された。予測構造評価方法の一つは、予測構造と正解構造の重ね合わせを行い、対応残基が3Åより短い距離にある残基数を出力Ａ～Ｄについて積算すること（sum値）により行われた。22のタンパク質を構造ドメイン単位（全部で３４ドメイン）で眺めた結果によれば、構造予測コンテストCASP5/CAFASP3における上記構造認識部門において22本の配列各々に対して上位1個の予測を考慮した時、本手法のsum値は「５７７」であり、これは、配列情報を用いた他のいかなる手法よりも優れているものであった 。

また、ある閾値を設定してある入力（予測対象）配列に対する予測の成否を観測した場合でも、22本の配列各々に対して上位1個の予測を考慮した時本手法は、予測が成功したと判断される個数を出力Ａ～Ｄについて積算したもの（correct値）において、「９」と高く、配列情報を用いた他のいかなる手法よりも優れていることが示された。

請求の範囲

1． タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の２つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

を含むシステム。

2． 請求の範囲第１項に記載のシステムにより作成されたスコア行列を用いることを特徴とするタンパク質立体構造の予測システム。

3． コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムとして機能させるためのプログラムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の２つのプロファイル行列を用意する手段と、

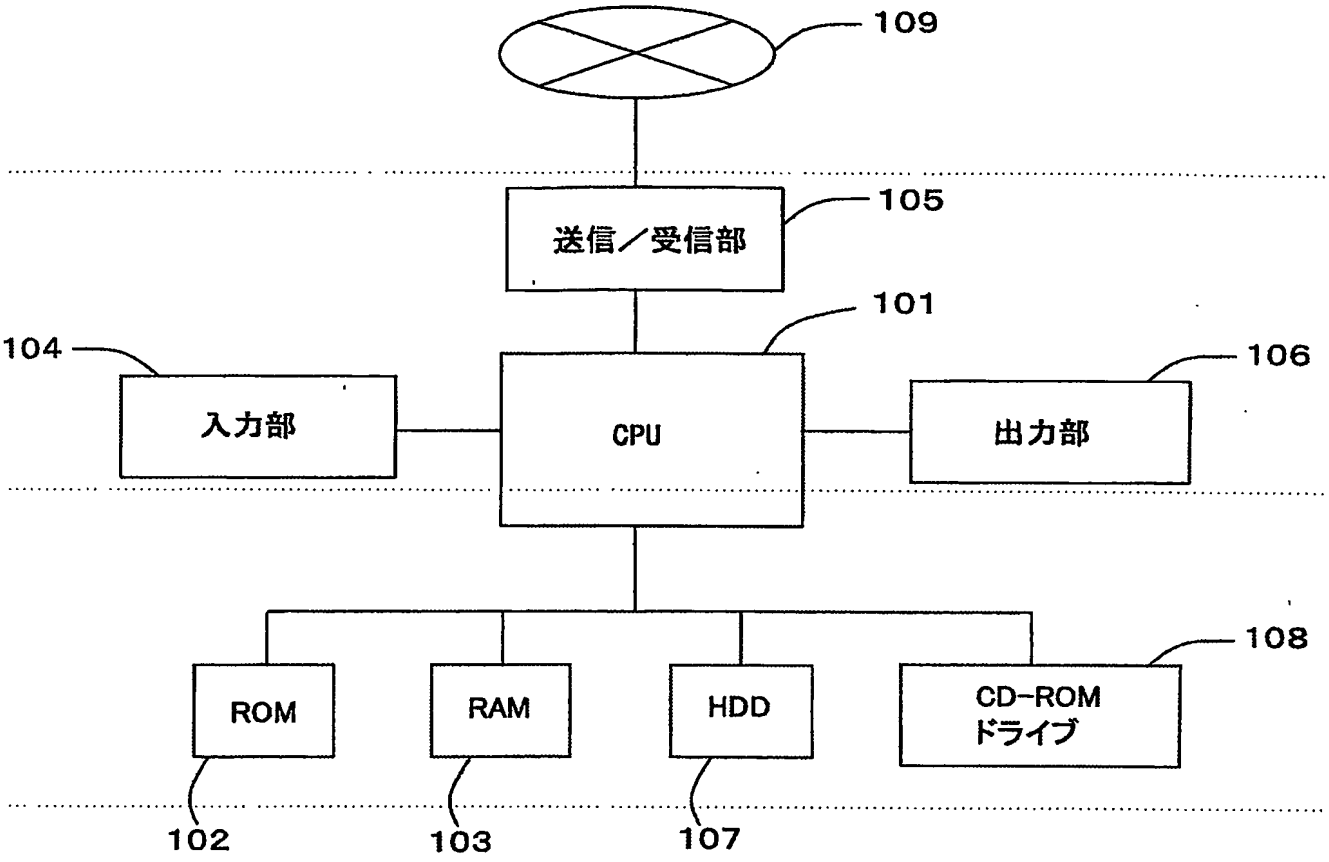
(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

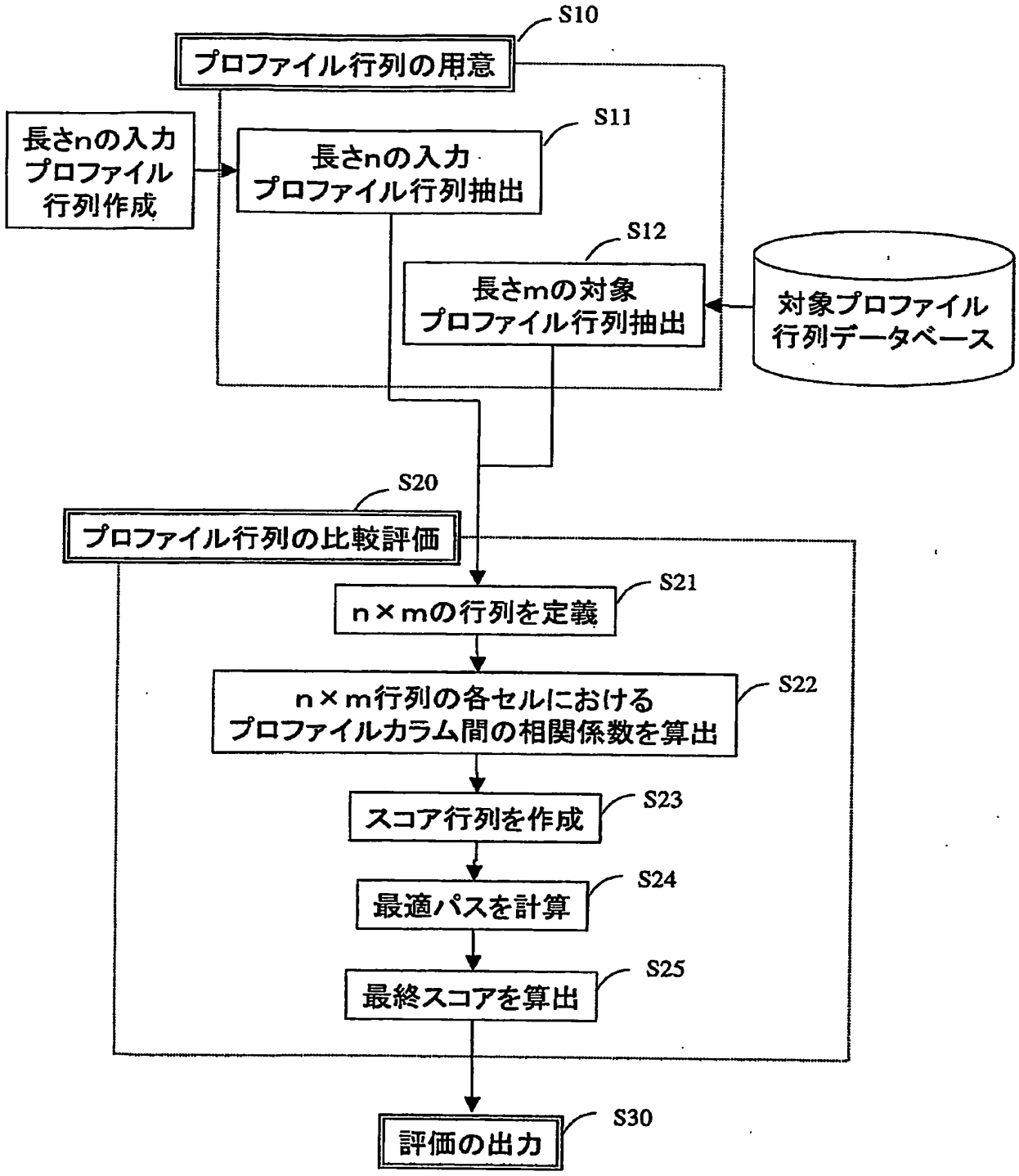
を含むプログラム。

4．請求の範囲第3項に記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

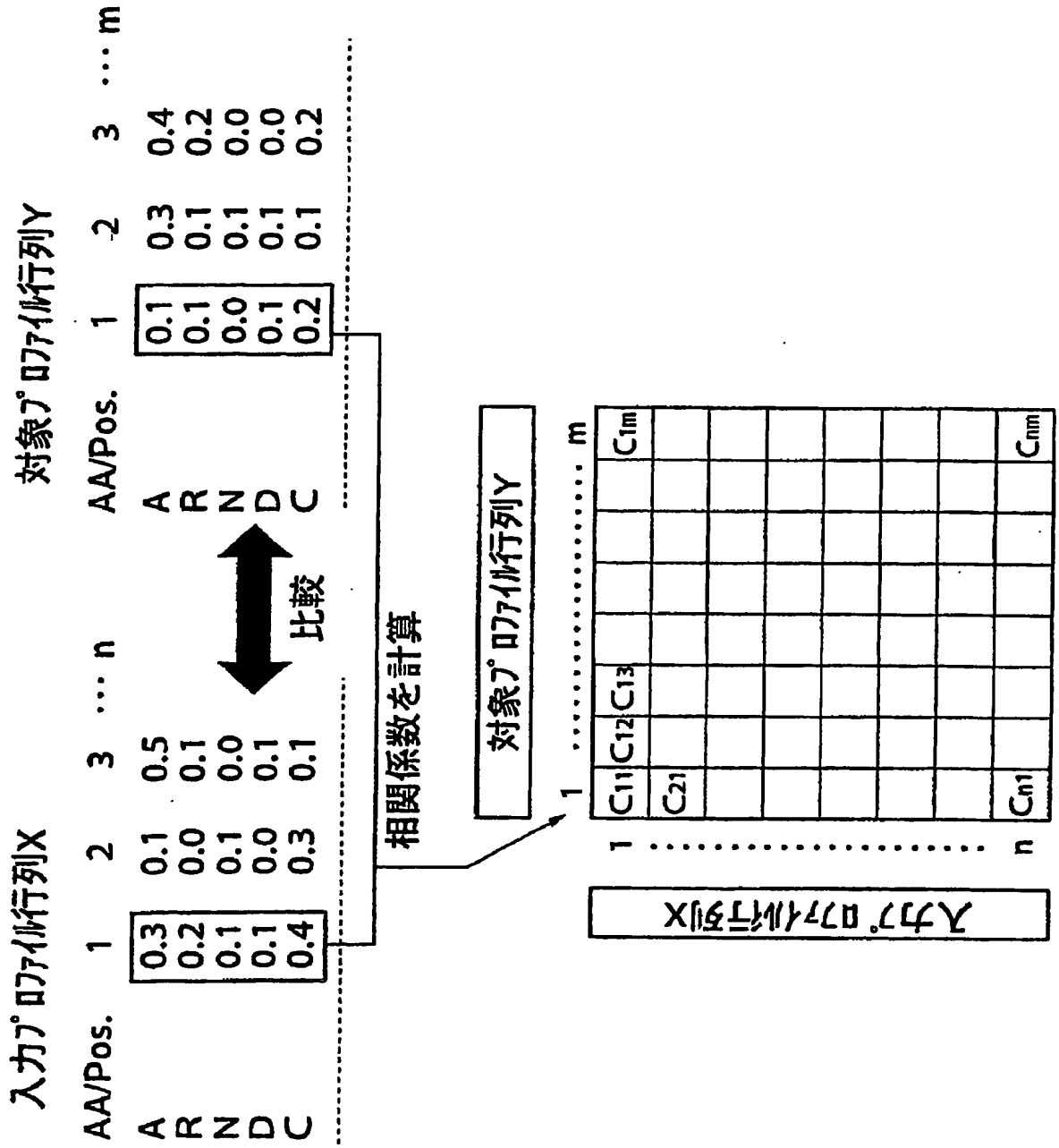
第１図


109
105
送信／受信部
101
104
入力部
CPU
106
出力部
108
ROM
RAM
HDD
CD-ROM
ドライブ
102
103
107

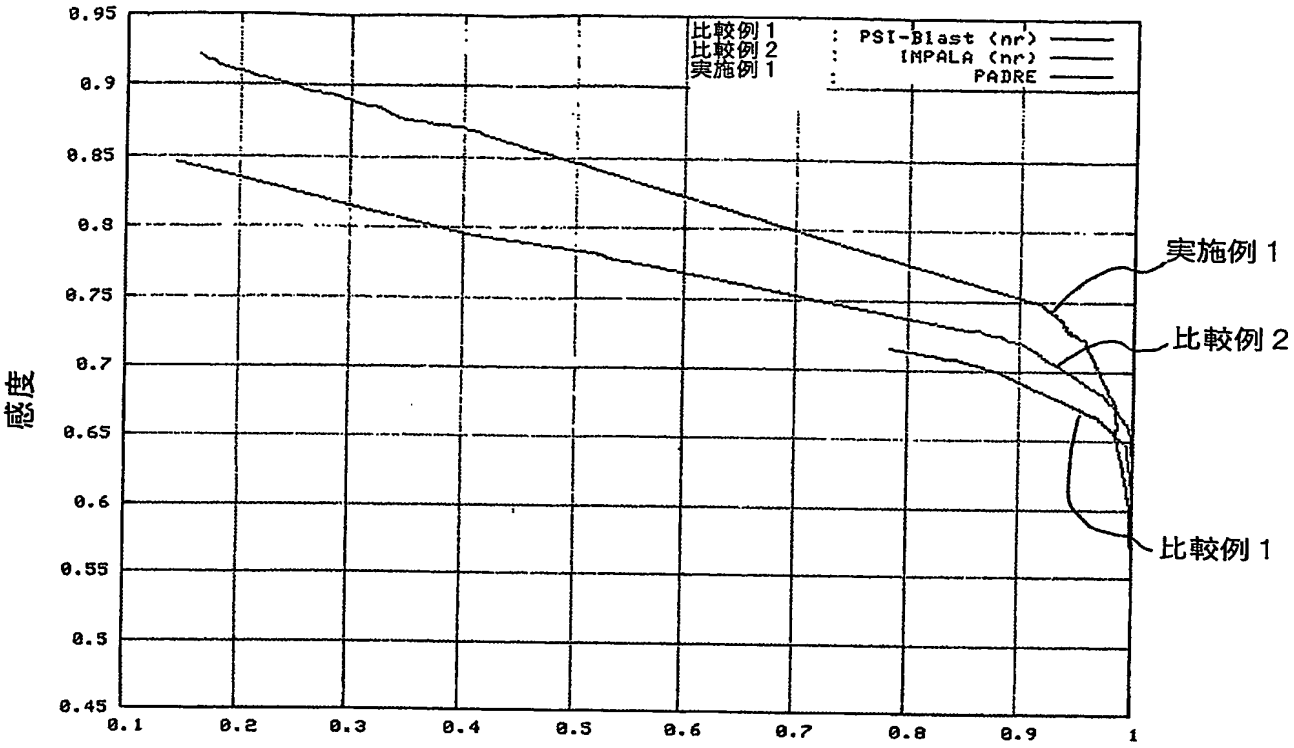
第２図


S10
プロファイル行列の用意
長さnの入力
プロファイル
行列作成
S11
長さnの入力
プロファイル行列抽出
S12
長さmの対象
プロファイル行列抽出
対象プロファイル
行列データベース
S20
プロファイル行列の比較評価
S21
n×mの行列を定義
S22
n×m行列の各セルにおける
プロファイルカラム間の相関係数を算出
S23
スコア行列を作成
S24
最適パスを計算
S25
最終スコアを算出
S30
評価の出力

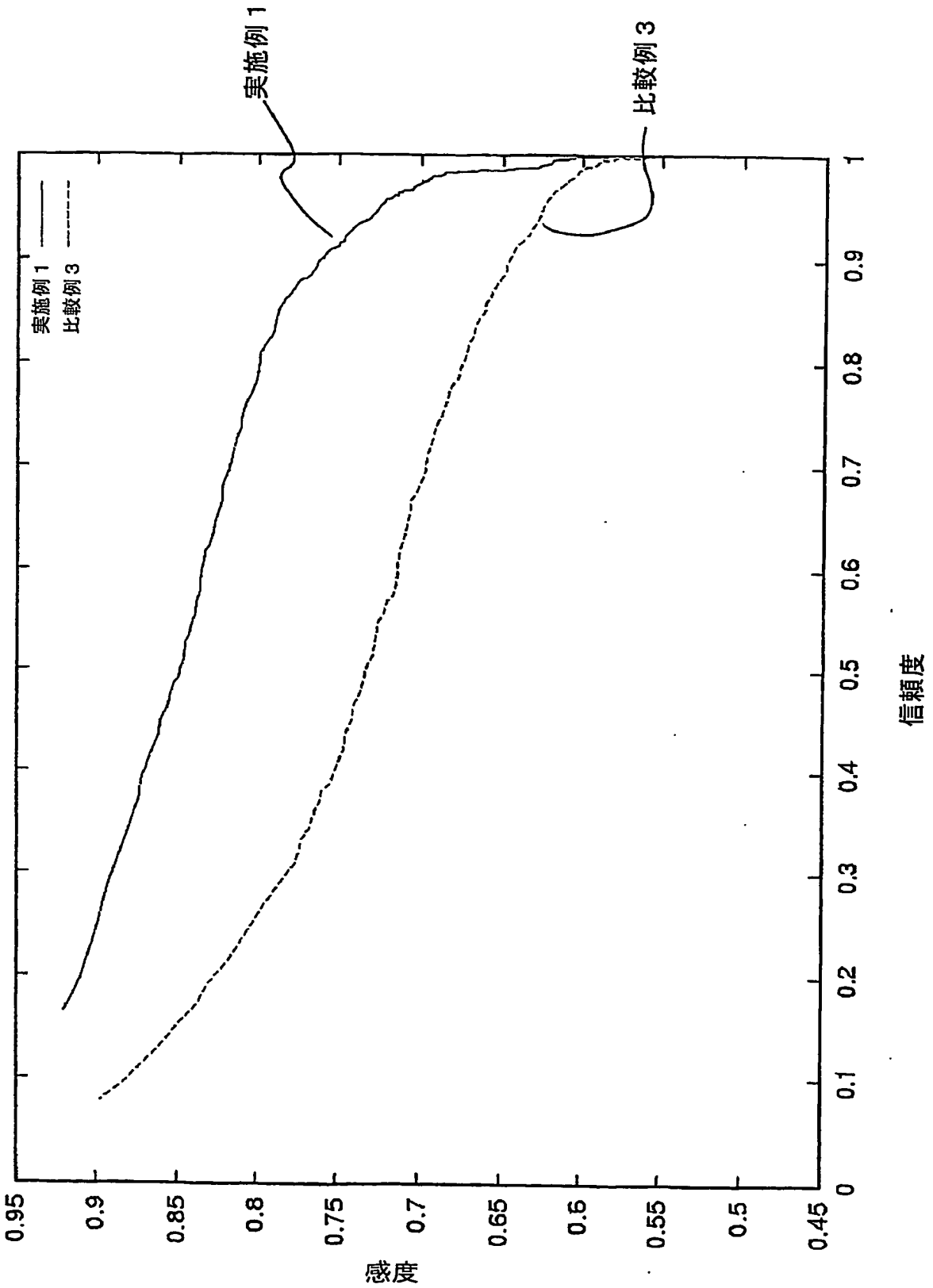
第３図


入力プロファイル行列X
対象プロファイル行列Y
AA/Pos.
A
R
N
D
C
1
2
3
… n
… m
0.3 0.2 0.1 0.1 0.4
0.1 0.0 0.1 0.0 0.3
0.5 0.1 0.0 0.1 0.1
0.1 0.1 0.0 0.1 0.2
0.4 0.2 0.0 0.0 0.2
比較
相関係数を計算
対象プロファイル行列Y
入力プロファイル行列X
C11 C12 C13 C1m
C21
Cn1 Cnm

第４図


比較例１
比較例２
実施例１
: PSI-Blast (nr)
: IMPALA (nr)
: PADRE
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
感度
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
信頼度
実施例１
比較例２
比較例１

第 5 図


実施例 1
比較例 3
実施例 1
比較例 3
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
感度
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
信頼度

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP03/16982

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F19/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Toroku Jitsuyo Shinan Koho	1994-2004
Kokai Jitsuyo Shinan Koho	1971-2004	Jitsuyo Shinan Toroku Koho	1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	YONA, G. LEVITT, M. 'Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based Information Theory', Journal of Molecular Biology, 01 February, 2002 (01.02.02), Vol.315, Issue 5, pages 1257 to 1275; especially, pages 1259 to 1260 [on line] [retrieved on 29 January, 2004 (29.01.04)], Retrieved from: <URL=http://www.cs.cornell.edu/golan/Papers/jmb02.pdf>	1-4
A	RYCHLEWSKI, L. et al., 'Comparison of sequence profiles. Strategies for structural predictions using sequence information', Protein Science, February 2000, Vol.9, Issue 2, pages 232 to 241 [on line] [retrieved on 29 January, 2004 (29.01.04)], retrieved from <http://www.proteinscience.org/cgi/reprint/9/2/232.pdf>	1-4

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 30 January, 2004 (30.01.04)	Date of mailing of the international search report 10 February, 2004 (10.02.04)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP03/16982

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2002-358309 A (Hitachi Software Engineering Co., Ltd.), 13 December, 2002 (13.12.02), & US 2002/184201 A1	1-4

国際調査報告

国際出願番号　PCT／JP03／16982

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl'　G06F19／00

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl'　G06F19／00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2004年
日本国登録実用新案公報	1994－2004年
日本国実用新案登録公報	1996－2004年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル（JOIS），WPI，INSPEC（DIALOG）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	YONA, G. LEVITT, M.'Within the Twilight Zone:A Sensitive Profile-Profile Comparison Tool Based Information Theory', Journal of Molecular Biology, 1 February 2002, Vol. 315, Issue 5, p. 1257-1275, especially p. 1259-1260 [on line] [retrieved on 29 January 2004], Retrieved from: <URL=http://www.cs.cornell.edu/golan/Papers/jmb02.pdf>	1－4

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日 30. 01. 2004	国際調査報告の発送日 10.2.2004
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 髙瀬　勤　5M　9069 電話番号　03－3581－1101　内線　3597

様式PCT／ISA／210（第2ページ）（1998年7月）

国際調査報告

国際出願番号 PCT／JP03／16982

C（続き）. 関連すると認められる文献

引用文献のカテゴリー＊	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	RYCHLEWSKI, L. et al. 'Comparison of sequence profiles. Strategies for structural predictions using sequence information', Protein Science, February 2000, Vol. 9. Issue 2, p. 232-241[on line][retrieved on 29 January 2004], retrieved from<http://www.proteinscience.org/cgi/reprint/9/2/232.pdf>	1－4
A	JP 2002-358309 A(日立ソフトウェアエンジニアリング)2002.12.13 &US 2002/184201 A1	1－4